Assignment 3: Rocket Fuel Case Analysis

Qiutong Shi CA Fri 8:30 – 11:30

1.1 share of users allocated to the control group

Using dplyr, the share of users allocated to the control group is 23524, 4% of the total number of users.

```
> ds %>%
+   group_by(test) %>%
+   summarise(percent = 100 * n() / nrow(ds) )
# A tibble: 2 × 2
   test percent
  <int>   <dbl>
1     0    4.00
2     1   96.0
> table(ds$test)

     0      1
 23524 564577
```

1.2 test for whether or not the campaign properly randomized consumers into the test and control group?

The idea behind the test randomization is to make sure seeing public service ad vs. real ad should not affect the number of impressions a user is served. If both groups have very different average of ad impressions seen, then it is possible that the number of impressions is behind the relationship between seeing an ad and purchasing.

This test will compare the average number of ad impression in the test group and control group by running t-test and checking p-value whether the two groups are significantly different.

1.2 test for whether or not the campaign properly randomized consumers into the test and control group?

I use split() to divide control/test group and corresponding impressions. Split() will return two elements. Then I use the t.test() to calculate the p-values.

```
> options(digits=15)#avoid R's automatic rounding by displaying more digitis
> ave<-split(ds$tot_impr, ds$test)
> t.test(ave[[1]],ave[[2]], alternative = "two.sided")

        Welch Two Sample t-test

data:  ave[[1]] and ave[[2]]
t = -0.2179969144924, df = 25607.80139496, p-value = 0.827433252496
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.621728606219399  0.497273522634920
sample estimates:
       mean of x        mean of y
24.7611375616392 24.8233651034314
```

The p-value is larger than 0.8, meaning that I cannot reject my null hypothesis. The randomization is proper such that there is no statistical difference between the mean of control and test group.

1.3 Why is randomization important?

Randomization is important in making sure there is no other unknown factor except the one we are testing that can influence the dependent variable. In rocketfuel case, Rocketfuel would want to prove that their ad is the major factor driving purchase decision, not other characteristics of consumers.

2. Was the campaign effective in increasing conversion rates?

The campaign was effective in increasing conversion rate. The conversion rate of test group is 0.0256, of control group is 0.0179. The one-tailed t-test has a p value equal to 1, failing to reject the null hypothesis that conversion rate of test group is larger than that of the control group.

```
> #one-sided test with null hypothesis that conversion of test is larger than control
> t.test(con[[2]], con[[1]],alternative = "less")

        Welch Two Sample t-test

data:  con[[2]] and con[[1]]
t = 8.657162314552, df = 26384.19095509, p-value = 1
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
                -Inf 0.00915406420888383
sample estimates:
         mean of x          mean of y
0.0255465596366837 0.0178541064444822
```

3.1 How much more money did TaskBella make by running the ad campaign?

I would want to compare the incremental revenue between exposed and non-exposed customers. Since there is no nonexposed customer in the data, I assume seeing 1 ad has very little impact on purchase decision. Therefore, I would compare those who saw 1 ad vs. those who saw more. If 1 ad has some impact, this will be an underestimate.

We learned that the incremental revenue is $556733.82.

```
> exp<-split(ds$converted, ds$tot_impr==1)
> #conversion rates
> con_exp = mean(exp[[1]])
> con_nexp = mean(exp[[2]])
> incremental_con = con_exp-con_nexp
> #count of converted given imp>1
> con_count_exp = length(exp[[1]])
> incremental_revenue = con_count_exp * incremental_con  * 40
> round(incremental_revenue,digits=2)
[1] 556733.82
```

3.2 What was the cost of the campaign?

The cost of the campaign is $131374.64, given information from the case.

Cost = cost of impression per thousand * number of impressions in thousands

```
> #cost of campaign
> #average CPM = $9
> cost = (14597182/1000)*9
> round(cost,digits=2)
[1] 131374.64
```

3.3 Calculate the ROI (i.e. Return on Investment) from the campaign.

The cost of the campaign is 323.78%, given results from 3.1 and 3.2 using the non-exposed vs. exposed method.

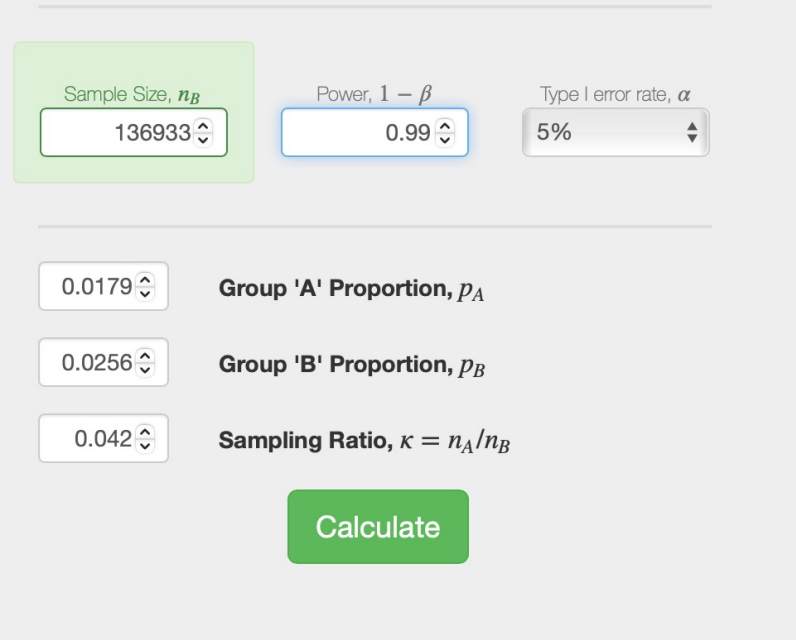
ROI = (incremental revenue – additional cost)/additional cost

```
> #ROI
> ROI = (incremental_revenue - cost)/cost * 100
> round(ROI, digits=2)
[1] 323.78
```

3.4 Was it worthwhile to use a control group? Could it have been smaller? Why or why not?

A control group was worthwhile in this case to see that the campaign directly helped the conversion rate, not another third factor.

Using the conversion rate 0.0256 of test group, 0.0179 of control group, with 0.4/0.96 = 0.042 sampling ratio, sensitivity of 0.99.



The smallest control group size would be 0.04/0.96*136933 = 5705. Currently, there are 23524 people in the control group. The control group could have been smaller by 23524-5705 = 17819.

4.1 Create and attach as exhibit (i.e. graph) a bar chart of conversion rates as a function of the number of ads displayed to consumers.

In order to produce the graph using ggplot, I created bins (0-9, 10-19,…,200+) as categorical variable and manually assign each variable to its bin by checking if it falls within the bin interval. I also create 2 bar group test and control. Then I use split() to split converted by myBins and test, and manually repeatedly assign conversion rate by myBins and test. myBins, conversion rate, and group are all added in the dataframe.

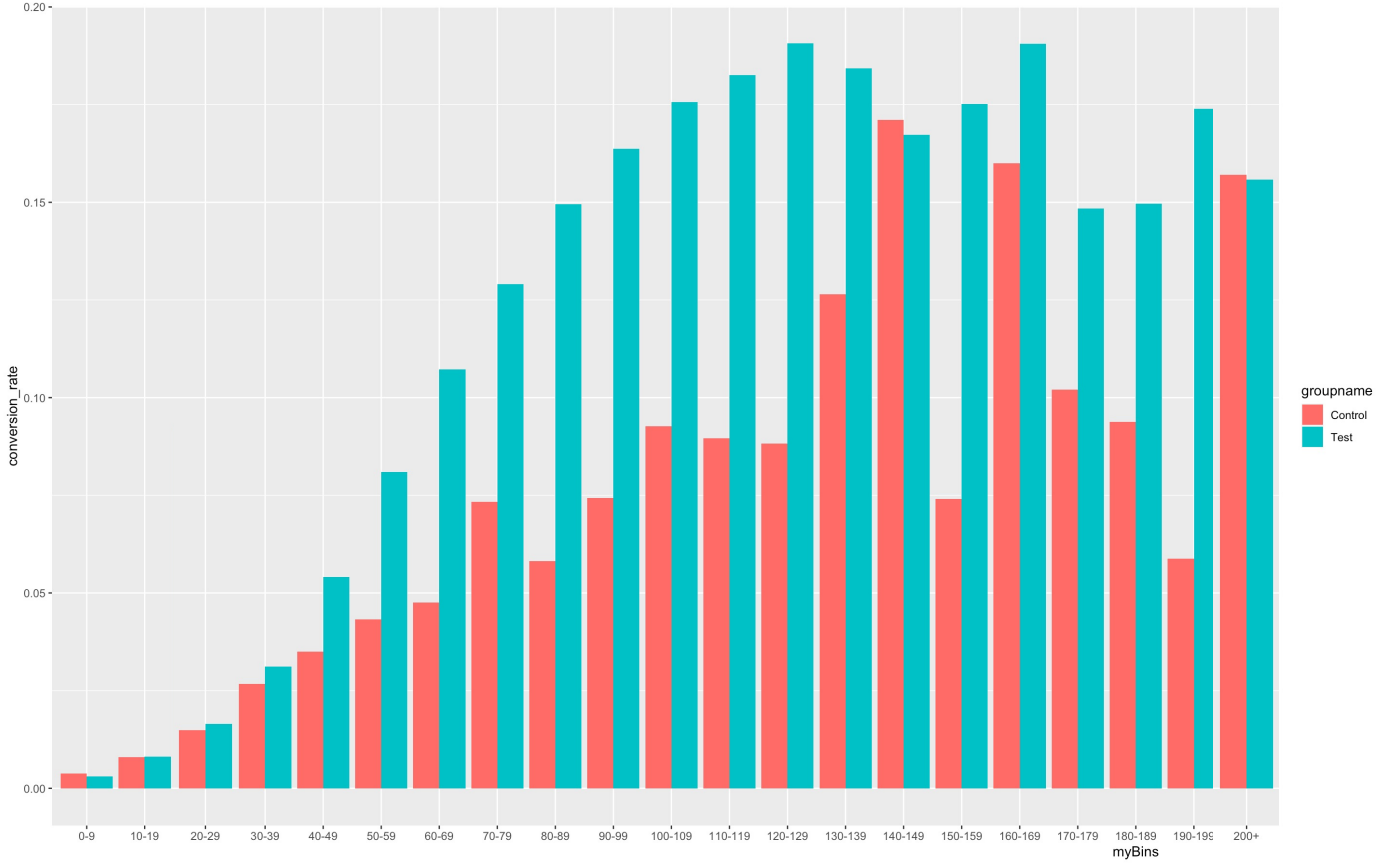
Refer to parts of the code on the next slide. The codes displayed here are partial because they are not tidy and probably have a large runtime because they are IFELSE inside IFELSE. One possible improvement I can think of is to use a loop and multiple conditionals inside to popularize the three new columns. Unfortunately, I always got error when trying it.

4.1 Create and attach as exhibit (i.e. graph) a bar chart of conversion rates as a function of the number of ads displayed to consumers.

```
# library
library(ggplot2)
#create and assign bins, groupname,
ds$myBins <- ifelse(ds$tot_impr>=0&ds$tot_impr<=9,'0-9',
             ifelse(ds$tot_impr>=10&ds$tot_impr<=19,'10-19',
             ifelse(ds$tot_impr>=20&ds$tot_impr<=29,'20-29',
             ifelse(ds$tot_impr>=30&ds$tot_impr<=39,'30-39',
             ifelse(ds$tot_impr>=40&ds$tot_impr<=49,'40-49',
             ifelse(ds$tot_impr>=50&ds$tot_impr<=59,'50-59',
```

```
ds$groupname<-ifelse(ds$test==0,'Control','Test')
#assign conversion rate based on group of different bin
bin_con<-split(ds$converted, list(ds$myBins,ds$test))
ds$conversion_rate <- ifelse(ds$test==0&ds$myBins=='0-9',mean(bin_con[["0-9.0"]]),
             ifelse(ds$test==0&ds$myBins=='10-19',mean(bin_con[['10-19.0']]),
             ifelse(ds$test==0&ds$myBins=='20-29',mean(bin_con[['20-29.0']]),
             ifelse(ds$test==0&ds$myBins=='30-39',mean(bin_con[['30-39.0']]),
             ifelse(ds$test==0&ds$myBins=='40-49',mean(bin_con[['40-49.0']]),
```

4.1 Create and attach as exhibit (i.e. graph) a bar chart of conversion rates as a function of the number of ads displayed to consumers.




```
ggplot(ds,
aes(fill=groupnam
e,
y=conversion_rate
, x=myBins)) +
geom_bar(positio
n="dodge",
stat="identity")
```

4.2 Is there a frequency effect to advertising, i.e. does showing more ads increase the probability of conversion?

Yes, there is a frequency effect to advertising. According to the graph, as the number of total impressions get larger, the difference between bars of the control group and test group within the same bin generally get larger at the same time.